Juniper Networks | Driven by Experience™

NETWORKING THE AI DATA CENTER

Juniper is powering the AI revolution with innovative networking technologies that speed data transfer, provide lossless transmission, and enhance congestion control.

White Paper

Table of Contents

Introduction

Recent advances in generative AI have captured the imagination of hundreds of millions of people around the world. These powerful AI models are changing the way many job functions are performed, including software development, graphic design, content marketing, and others.

Developing an AI model is a lengthy and challenging process built upon three foundational phases. The first phase, data preparation, involves gathering and curating datasets so that they can be fed into the AI model. The second phase, AI training, is the process of teaching an AI model to perform a specific task by exposing it to large amounts of data. During the training phase, an AI deep neural network learns patterns and relationships within the training data to develop virtual synapses to mimic intelligence. Once the model is trained, it is transitioned to the AI inference phase where it operates in a real-world environment to make predictions or decisions based on new, unseen data.

The training phase is a deep learning, iterative process where the AI model learns from the gathered data to refine its parameters while the inference phase focuses on applying the learned knowledge to new inputs. Applications like **ChatGPT**, Meta LLaMA, and Google LaMDA are examples of Large Language Models (LLM) using the inference phase to answer questions, write college essays, code games, and compose hit songs.

AI efficacy is the degree to which an AI model performs as intended, while AI training efficiency is the amount of time and resources required to train an AI model to achieve its desired level of performance. For example, pre-training times for Meta's Llama 2 model ranged from 184K GPU-hours for the 7 billion-parameter model to 1.7M GPU-hours for the 70 billion-parameter model. Any inefficiency that lowers the GPU processing speed would increase the model training time. Likewise, increasing or decreasing the number of parameters or GPUs will impact training time. Adding compute is the logical antidote to reduce training time, but GPUs are expensive and account for up to 80% of AI training costs. A Juniper team researching these costs estimates that an AI training server powered by 8 GPUs can cost in excess of $400,000.

Consequently, the efficiency and utilization of these expensive compute resources is critical to containing costs. Data center networks can be bottlenecks for that efficiency because many hundreds or thousands—even tens of thousands—of GPU servers must be connected to train large models. AI training efficiency is measured in job completion time which starts with the distribution of workloads to GPU clusters and ends when the last workload is fully processed and merged.

Minimizing or eliminating tail latency, a condition where outlier AI workloads slow down the completion of the entire AI job, is key to optimizing job completion time and maximizing the return on GPU investment.

This paper looks deeper into the challenges and requirements associated with AI infrastructure data center networks and Juniper Networks' recommendations for their design, deployment, and operation. Juniper's approach to networking the AI data center is based on our corporate DNA and obsession with massively scalable network performance, industry-standard openness and innovation, and experience-first operations.

AI Data Center Networking Overview

AI training is the most technologically disruptive and challenging part of the overall AI process, particularly for complex deep-learning models requiring massive amounts of data and distributed processing to achieve optimal performance. For example, training a state-of-the-art image recognition model can require millions of labeled images. Once preprocessed and curated, the dataset is fed into the AI model. Many modern-day deep learning models have billions of parameters, some even have trillions. The parameters are also known as weights that get adjusted to make inference more accurate over many training iterations.

AI inference clusters are connected to front-end networking fabrics that connect to the outside world to support inference requests from users, IoT devices, and more. While compute-intensive, front-end network traffic patterns are typical to other cloud computing models, achieving GPU efficiency with classic data center networking designs and technologies. In contrast, distributed AI model training requires additional considerations to optimize the overall compute across the network where GPUs depend on one another to exchange and compute parameter weights. If the network is a bottleneck that delays job completion, expensive compute time is wasted, and training becomes network-bound instead of compute-bound. Hence, GPUs are connected together with a high-performance dedicated fabric to speed up training. This dedicated fabric is called a back-end fabric.

Per figure 1, back-end network fabrics support both GPU training clusters and AI storage systems. Implemented as separate back-end fabrics, compute training clusters and storage networks both provide high performance, low latency networking for each service respectively. While smaller AI systems may not need dedicated back-end fabrics, the majority of AI deployments larger than a couple of racks of state-of-the-art GPUs do require them to optimize training efficiency and storage speed.

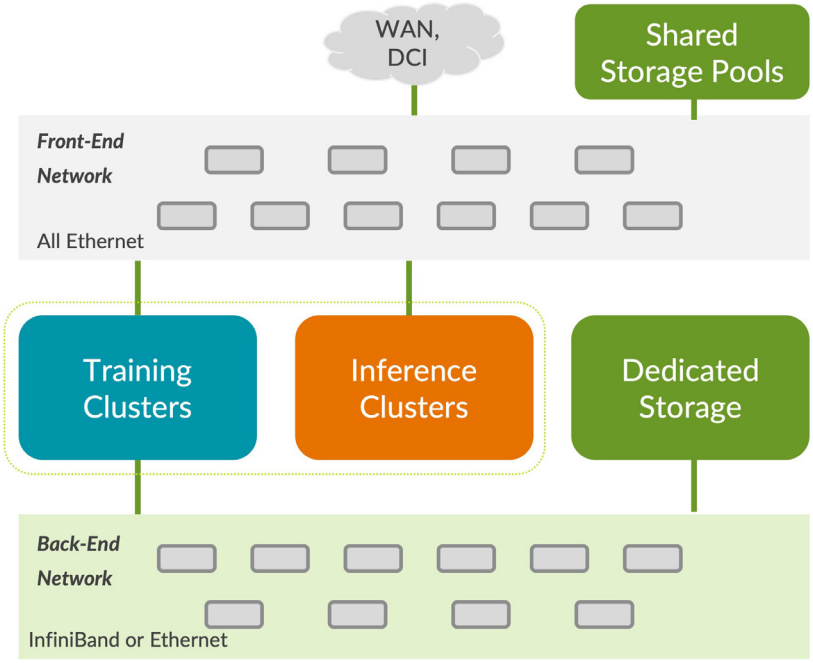

Figure 1: AI front-end and back-end networking

The GPU training back-end fabric and the workloads running over them present more like a high bandwidth "echo chamber" as training techniques iterate on the model. Contemporary GPUs are connected to the fabric at 400 Gbps. And popular servers have eight such GPUs and corresponding network interfaces. The fabric experiences unique traffic patterns of long-lived, low entropy, bandwidth-intensive flows, requiring congestion avoidance. As AI/ML cluster costs can go up to hundreds of millions of dollars for the largest clusters, networking is a key enabler of GPU efficiency.

To meet these challenges in early High-Performance Computing (HPC) and AI training networks, Remote Direct Memory Access (RDMA) became a key component of InfiniBand, a high-speed, low-latency proprietary networking technology that initially gained popularity for its fast and efficient communication between servers and storage systems. The InfiniBand technology, however, is solely sourced through Nvidia. Today, the open alternative to InfiniBand is Ethernet, which is gaining significant traction in the market for modern AI/ML network infrastructure and is expected to be the dominant technology of choice over InfiniBand. Ethernet is gaining fast adoption for multiple reasons, but operations and cost stand apart. The pool of talent that can build and operate an Ethernet versus a proprietary InfiniBand network is massive, and the tools available to manage such networks are also significant.

RDMA over Converged Ethernet (RoCE) has emerged as a popular option to support new storage, HPC, and AI network fabrics. This enables data center operators to benefit from the open standard Ethernet protocol. Ethernet evolves quickly in connectivity speeds, at lower costs, and is supported by a more diverse vendor landscape versus proprietary approaches.

With the increase in AI modeling across private, public, and hybrid clouds, and recent InfiniBand supply chain delays, the demand for more open and converged Ethernet solutions to improve flexibility, economics, and investment protection is increasing significantly. Already expensive, supply contention has further widened the financial gap between InfiniBand fabrics and open, Ethernet alternatives.

Today, RoCE is a key enabler allowing AI model operations and infrastructure teams to extend RDMA functionality to open, high-speed 400GbE and 800GbE fabrics with line of sight to 1.6 Terabit Ethernet in the future.

Training Network Fabric Traffic Flows and Bottlenecks

Traditional spine-leaf fabrics used to support cloud data centers and their non-AI applications are inadequate for RoCE traffic patterns during training. AI training increases the potential for collisions and GPU contention that can congest the network switches and links. Like cars entering a toll booth plaza, throughput is smooth when traffic is distributed evenly across each toll booth. But AI training uses a small number of very large, long-lived flows (elephant flows) that act like trucks entering the same toll plaza. As elephant flows converge at vulnerable points in the fabric, congestion can occur leading to collisions and packet drops resulting in retransmissions that ultimately increase the training job completion time.

Other than the challenges of incast, where multiple senders may target a single receiver, data center network fabrics have two primary network congestion points: the leaf-to-spine uplinks and the spine-to-leaf downlinks.

Leaf-to-Spine Uplink Congestion

Capable of and likely running multiple flows simultaneously, GPUs are powerful enough to consume full network interface capacity, up to 800 Gbps today. As flows from GPUs are load balanced at leaf switches across dozens of uplinks towards spine switches, imbalanced flow assignment can congest the uplinks, ultimately impacting job completion time. Even without oversubscription, equal cost load balancing of flows is based on hashing to choose a spine-facing link. Multiple flows can end up congesting a link while other links remain underutilized.

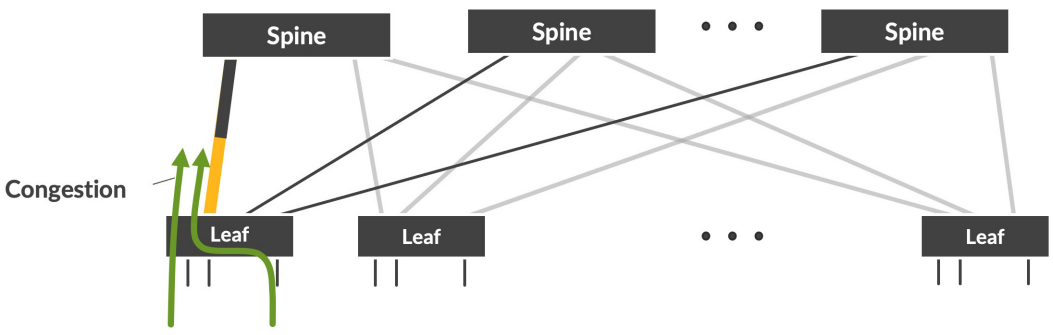

Figure 2: Congestion at the leaf uplink

Spine-to-Leaf Downlink Congestion

Inversely, traffic from several leaf switches through the spine may be directed to the same leaf. With finite capacity and possibly a single spine-to-leaf downlink, flows can oversubscribe the interface resulting in buffer exhaustion, congestion, and drops. Note that load balancing decisions are made independently by ingress leaf switches, making this congestion scenario particularly difficult to address.

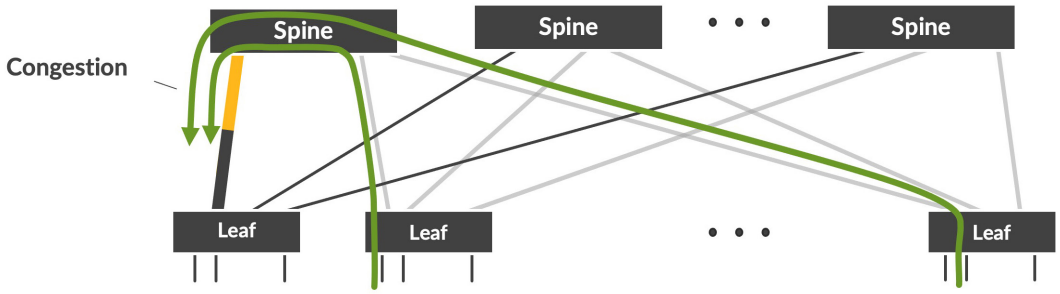

Figure 3: Congestion at the spine

Imperfect load balancing reduces the effective transmission rate through the fabric. The effective transmission rate is a function of two variables:

- Number of flows–transmission rate increases with more flows
- Number of paths–transmission rate decreases with more paths

A relatively small number of large flows, a common condition with AI workloads, presents a difficult problem for traditional network designs employing per-flow load balancing. In the earlier toll booth metaphor, a small stream of cars, buses, and trucks entering a toll booth plaza will result in highly variable transit rates through each toll booth. As the number of vehicles increases, the load balancing across toll booths will statistically improve to normalize transit rates and increase overall efficiency.

Load-balancing efficiency and congestion-management protocols are important to serve back-end network fabrics supporting AI training workloads. Network bottlenecks and inefficiencies that go undetected or unresolved can have costly impacts on the AI infrastructure. To restate these concepts, if the network is a bottleneck for training, expensive compute time is wasted, and training becomes network bound instead of compute bound.

While proprietary, scheduled Ethernet fabric solutions exist today that improve load balancing, they introduce operational and visibility challenges and suffer from the same vendor lock-in dependency as InfiniBand fabrics. The ideal approach to solve AI networking challenges is based on open standard and interoperable Ethernet fabrics, with a focus on improving networking operations that meet the specific demands of different AI workload types.

Ethernet is an open, versatile data center networking technology with the performance, scale, and low latency needed for AI training and inference. Innovating at the pace of an entire industry rather than only a single vendor, Ethernet has evolved to keep pace with the networking demands of advanced computing environments. Available in fixed form factors or large chassis switches for multiplanar, multistage Clos or flatter, high-radix spine topologies, Ethernet is the most cost-effective and flexible data center technology available. As a converged technology, Ethernet fabrics facilitate multivendor integration and operations with flexible design options to match performance, resiliency, and cost goals for AI data center's back-end networking and extended AI infrastructures.

Juniper Networks' Approach to AI Networking

Juniper's AI data center solution vision builds upon our decades of networking experience and our AI-Native Networking Platform to deliver AI infrastructure that integrates with existing data centers using technologies that are open, flexible, fast, and simple to manage. Juniper helps customers deploy high-capacity, scalable, non-blocking fabrics that deliver the highest AI performance, the fastest job completion time, and the most efficient GPU utilization to improve economics leveraging open, standards-based Ethernet solutions. This approach offers three key benefits:

- **Operations-first approach** provides simple and seamless operator experiences that save time and money without vendor lock-in
- **Standards-based Ethernet solution** ensures feature velocity and cost savings
- **End-to-end solutions** build high-performing, scalable AI data centers with flexibility and ease

Networking Performance to Optimize GPU Efficiency

GPU efficiency is a quotient of fabric efficiency, capacity, scale, and performance. All are needed to effectively scale diverse AI models and ensure headroom for growth. The network is critical to both maximizing GPU utilization and optimizing the overall economics of an AI solution.

High-Capacity, Lossless AI Back-end Fabric Design

The combination of parallel processing and inter-GPU traffic creates long-lived and bandwidth-intensive traffic patterns that can congest the network. While a chosen AI/ML software framework (e.g., TensorFlow, PyTorch) may support various data center topologies, an any-to-any non-blocking Clos fabric design is recommended as the most versatile topology to optimize any training framework. To reduce the inefficiencies of speed shifts, AI training fabrics are built using a consistent networking speed (400Gbps shifting to 800Gbps) that extends from the NIC to the leaf and through the spine.

Building on that standard topology for scale and capacity, Juniper recommends a 2-layer, 3-stage, non-blocking Ethernet fabric with the flexibility and ease to extend to a 3-layer, 5-stage, non-blocking Ethernet fabric based on model size and GPU scale. As the figurative and literal backbone, the Juniper design recommends Juniper Networks® PTX10000 line

of routers based on Juniper Express Silicon for the spine and/or super spine with Juniper Network QFX5200 line of Switches based on Broadcom's Tomahawk ASICs as leaf switches to provide connectivity toward the servers. In state-of-the-art GPU clusters of 1024 or less, the same QFX Series switch, such as the QFX5240 with 64 x 800GbE ports, can serve as both a leaf and spine. to improve sparing. For clusters needing to grow larger, the PTX Series modular chassis provide a flexible path to growth.

As a high-radix spine switch, the PTX10000 line of Routers scales up to 460.8Tbps today with industry-leading 800GbE Ethernet density of 576 x 800GbE ports across up to 16 line cards in the largest chassis. A large port count, high-radix, spine device like this one allows for far more GPU compute nodes in a flatter 2-layer fabric before needing to add a third super-spine layer. For example, over 18,000 GPUs, each with 800GbE interfaces, can be connected into a single 2-layer Clos fabric.

For such massively scalable, highly redundant AI fabrics, Juniper recommends QFX5230 or QFX5240 leaf switches to extend server connectivity to the GPU compute access at up to 400Gbps or 800Gbps, increasing overall fabric capacity with high-speed uplinks to PTX spine switches (see Figure 4). In this scenario, leaf switches are assumed to be co-located next to the spine switches. Cabling options and optics are distance dependent but for cost considerations, inexpensive Active Optical Cables (AOC) or Active Electrical Cables (AEC) cables are recommended if distance allows. There is no strict latency variation and latency constraints across fabric links in the Juniper design.

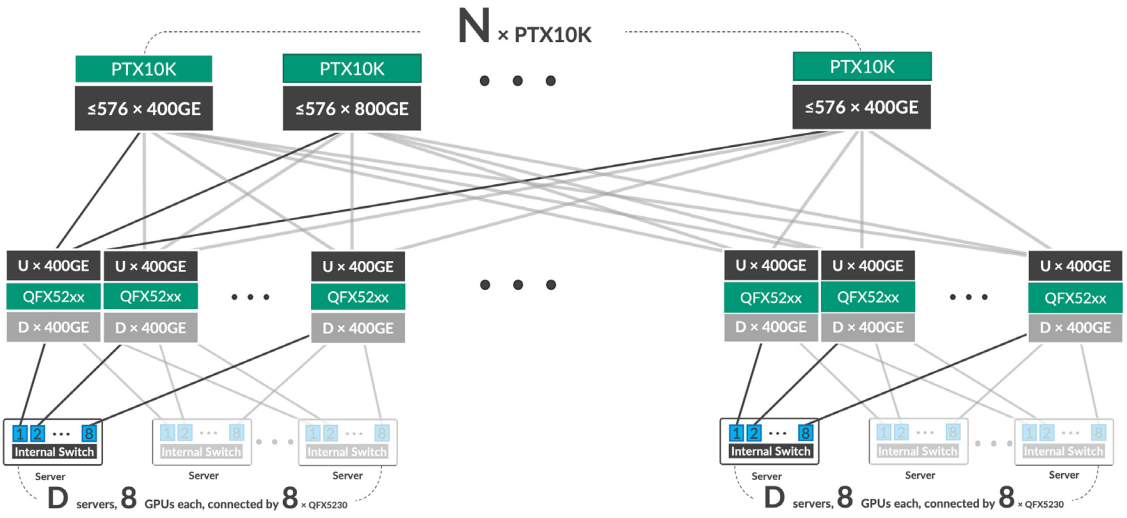

Figure 4: Large AI back-end network fabric topology

High Performance with ASIC Diversity

As a networking pioneer, Juniper ASICs have redefined the industry's benchmarks for packet processing, scale, and programmability. Building on our 400GbE leadership with the PTX Express 4 silicon, Juniper is leading the 800GbE evolution with Express 5, a feature-rich silicon designed for 800GbE AI data centers. Optimized for massive throughput and sustainable power efficiency, the Express 5 ASIC delivers high performance, 800GbE investment protection, and intelligent flow management to reduce flow-tail latency, and in turn, AI training job completion time.

Extending 800Gbps to the leaf and spine fixed form factor platforms, Juniper is incorporating the most advanced Broadcom silicon within our QFX Series of data center switches. By offering custom and merchant silicon diversity, customers maintain architectural choice and flexibility based upon design criteria, use case, feature set, and platform preference.

Customers who have invested in Juniper 400GbE fabrics are assured an easy migration to 800GbE and beyond.

Fabric Efficiency with Flow Control and Congestion Avoidance

Fabric capacity alone does not guarantee fabric or GPU efficiency. Design considerations can increase throughput, reduce congestion, and improve reliability that contribute to the coverall efficiency of the fabric. Considerations include but are not limited to:

- Properly-sized fabric interconnects with the optimal number of links for capacity, resiliency, and flow balancing efficiency
- Detection and correction of flow imbalances to avoid congestion and packet loss and ensure high-priority, lossless traffic is not impacted

While Juniper's AI networking fabric designs minimize the likelihood of congestion due to flow imbalances, congestion may still occur. To compensate, flow control mechanisms based on Explicit Congestion Notification (ECN), Data Center Quantized Congestion Notification (DCQCN), plus Priority-Based Flow Control resolve these conditions to ensure lossless transmission.

Load Balancing

Non-uniform load balancing over fabric links is the main cause of congestion in the AI back-end network fabric. Load balancing detects flow imbalances locally at the switch to redistribute flows more evenly across the fabric. While different methods detect, apply, and change hashing functions, link utilization and queue size averages must be periodically checked to identify and rebalance flow rate disparity, using periods of flow inactivity to eliminate packet re-ordering.

Explicit Congestion Notification

Juniper's AI networking fabric fully supports ECN marking to provide early indications of network congestion to applications. During periods of congestion, leaf and spine switches update ECN-capable packets to notify senders of the congestion condition. Receivers of the packets react to the congestion signals and request senders to slow down the transmission to avoid packet drops in transit.

If the end points do not react to the ECN congestion indication in time, Priority-Based Flow Control (PFC), explained in the next section, allows the fabric itself to pause transmission.

Priority-Based Flow Control

During periods of heavy traffic, switch buffers can fill up leading to silent packet drops and retransmissions that exacerbate congestion. PFC allows Ethernet receivers to share feedback with senders on buffer availability. During periods of congestion, leaf and spine switches may pause or throttle traffic on specific links to reduce congestion and avoid packet drops. PFC is an important technology across Juniper's AI networking fabric to provide lossless transmissions for specific traffic classes.

Open, Standards-Based Ethernet Scale and Performance

Massive AI datasets have driven the need for greater compute power, faster storage, and high-capacity, low-latency networking. As parallelism and hardware acceleration have modernized compute and storage systems, Ethernet has emerged as the open-standard solution of choice to handle the rigors of HPC and AI applications.

Next to IP, Ethernet is the world's most widely adopted networking technology. Ethernet has evolved to become faster, more reliable, and scalable, making it preferred for handling the high data throughput and low-latency requirements of AI applications. The progression to 800GbE and data center bridging (DCB) Ethernet enhancements enable high capacity, low latency, and lossless data transmission, making Ethernet fabrics highly desirable for high-priority and mission-critical AI traffic. Juniper's commitment to open-standard technologies like Ethernet helps drive a culture of collaboration, transparency, and shared knowledge, allowing a diverse community to work together to continuously improve and innovate. To formalize that collaboration, Juniper joined the Ultra Ethernet Consortium (UEC) to accelerate the development of a common, high-performance Ethernet architecture for multivendor AI networks.

While proprietary technologies like InfiniBand can bring advancements and innovation, they are expensive, charging premiums where competitive supply and demand markets can't regulate costs. Ethernet currently dominates data center networking, including front-end AI networks. With advancements like DCB and RoCEv2, Ethernet is quickly becoming the AI back-end fabric of choice.

Highly Automated to Bring Speed and Simplicity to AI

As the industry has converged on IP fabrics and Ethernet VPN–Virtual Extensible LAN (EVPN-VXLAN) fabric designs for data centers, Juniper has focused on delivering experience-first operations to the companies that design, deploy, and manage data centers. This approach extends to new AI data centers. This means automating fabric management with Juniper® Apstra, intent-based networking software that automates and validates the data center network life cycle from Day 0 through Day 2+. Apstra translates business intent and technical objectives to essential policy and device-specific configuration and resolves issues to assure compliance. Its multivendor automation capabilities provide an abstraction layer based on logical designs across vendors, empowering organizations to automate and manage their networks across virtually any data center location, vendor, and topology.

Apstra blueprints are repeatable and continuously validated, ensuring that configurations are deployed correctly from the first time and every time. Blueprints also remove human error by pre- and post-validation of intent, so everything always works as intended. The robust telemetry (intent-based analytics) and flow data capabilities in Apstra provide valuable insights to ensure optimal network performance, facilitate proactive troubleshooting, and avert outages with predictive insights.

In recent research from Forrester, Apstra has proven 320% typical ROI in traditional data center deployments, and we expect similar results for new AI data centers using Juniper's AI networking fabric.

Rail-Optimized Design Ease with Apstra

With Apstra, designing for the Nvidia rail-optimized design is easy. In this design, each of the 8 GPU interfaces on the server is linked to a different leaf switch with logical connections called "rails" to other servers and GPUs connected to the same leaf switches. Allowing GPUs to write to buffers on non-local intermediate NICs, rail-optimized designs maximize performance while minimizing network interference between flows. Designing fabrics with these groups of 8 leafs and 8 server interfaces is possible using the flexible Apstra Logical Device and Rack components of design templating. Juniper has published many open-source examples with popular server types and rail-optimized groups of 8 leafs. Importing these into Apstra with Terraform is done in seconds. Importing these into Apstra with Terraform is done in seconds.

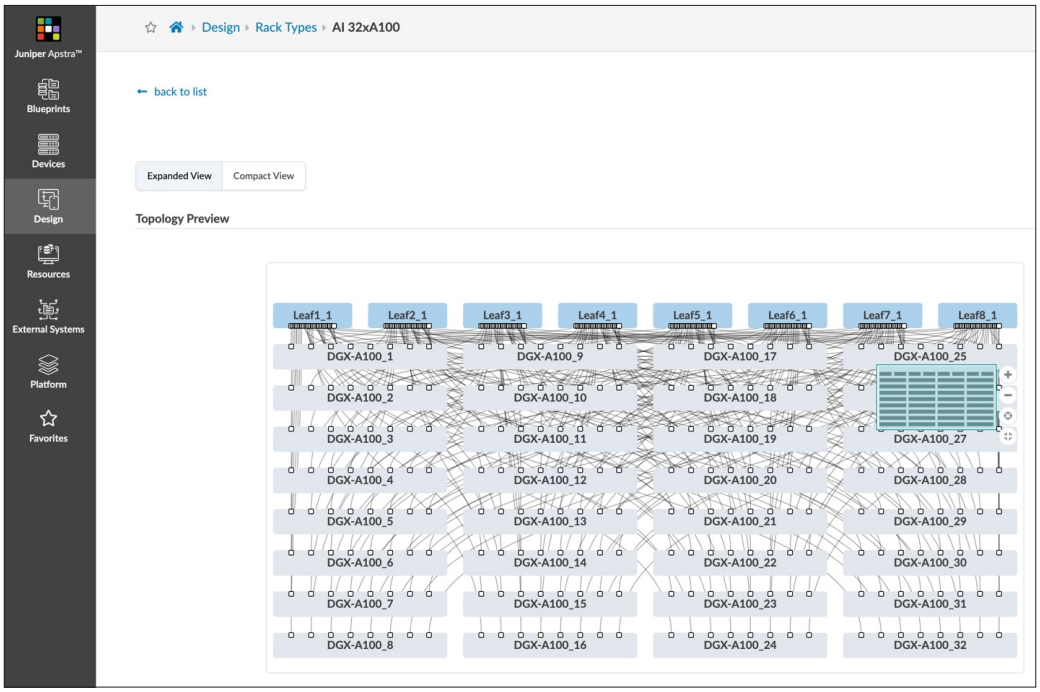

Figure 5: Apstra example of rail-optimized design

Apstra simplifies managing the back-end, front-end, and storage fabrics from a single pane of glass and common workflow while preserving customer choice of network hardware vendor between fabrics or network layers.

Summary

AI is now mainstream, but we are still in the early stages, scratching the surface of what is possible today and understanding how AI can and will impact us personally and professionally in the future.

Nearly every AI application has a data center behind it and these data center networks will remain a critical enabler in the decades ahead as we investigate AI boundaries. To meet the rigors of these demanding AI training environments, Juniper is delivering AI infrastructure solutions that deliver high performance to optimize GPU efficiency. Juniper solutions are built upon industry-standard Ethernet fabrics to optimize existing expertise while leveraging the vast Ethernet ecosystem, and they deliver experience-first operations to simplify and economize AI networking.

Juniper offers a diverse set of networking platforms, interface speeds, fixed form factors as well as modular Ethernet platforms for multistage Clos or flatter high-radix spine topologies. A market leader in 400GbE with 800GbE solutions available soon, we are committed to silicon diversity, using both merchant and custom silicon with options for shallow and deep buffers, giving customers choice and flexibility. Additionally, Apstra is the only multivendor, intent-based network automation and operations solution for efficient AI model training.

AI is an important inflection in computer and network communications, and Juniper is committed to driving innovative networking technologies that speed data transfer, provide lossless transmission, and enhance congestion control–critical aspects to powering the AI revolution.

To learn more about Juniper AI solutions, visit **Juniper.net** or read Juniper CEO Rami Rahim's **blog** on Juniper's AI-Native Networking Platform.

About Juniper Networks

At Juniper Networks, we are dedicated to dramatically simplifying network operations and driving superior experiences for end users. Our solutions deliver industry-leading insight, automation, security and AI to drive real business results. We believe that powering connections will bring us closer together while empowering us all to solve the world's greatest challenges of well-being, sustainability and equality.

JUNIPER NETWORKS | Driven by Experience™

APAC and EMEA Headquarters
Juniper Networks International B.V.
Boeing Avenue 240
1119 PZ Schiphol-Rijk
Amsterdam, The Netherlands
Phone: +31.207.125.700
Fax: +31.207.125.701

Corporate and Sales Headquarters
Juniper Networks, Inc.
1133 Innovation Way
Sunnyvale, CA 94089 USA
Phone: 888.JUNIPER (888.586.4737)
or +1.408.745.2000 | Fax: +1.408.745.2100
www.juniper.net

2000811-003-EN Jan 2024